

People, Places & Privacy

Using Finite State Machines to preserve privacy while data mining the cellular phone network

Author

Jonathan Reades, Bartlett School of Planning, University College London

Email:

j.reades@ucl.ac.uk

Address:

Planning Dept., 4th fl.,

The Bartlett School,

Wates House

22 Gordon St.

London WC1H 0QB

United Kingdom

Abstract

Keywords

Privacy, identity, mobile, cellular, phone network, finite state machine, locational data, spatio-temporal data, anonymization, LBS, GPS, tracking.

People, Places & Privacy

Using Finite State Machines to preserve privacy while mining data from cellular phone networks

Abstract

Keywords

Privacy, identity, mobile, cellular, phone network, finite state machine, locational data, spatio-temporal data, anonymization, LBS, GPS, tracking.

Introduction

In the past ten years there has been a proliferation of devices – handheld GPS units, mobile phones, short-range Radio Frequency Identification (RFID) tags, and video surveillance cameras, to name just a few – capable of sensing or tracking the movement of individuals through time and space. In the broader context, Lazer et al. [16] argue that the leveraging of these rich sources of behavioural data is ushering in a ‘computational social science’ markedly different in scale and scope from existing practices in fields such as sociology or planning. Researchers have already begun analysing data sets from cellular-network operators in Europe, America, and Africa with a view to understanding the links between such varying characteristics of social existence as individual mobility and commuting, public health, group dynamics, and deprivation, [1, 6, 9, 20] and the data underlying this research is truly massive: in one month, a single phone company’s call data records (CDRs) can easily exceed seven billion records!

However, the nature of the cellular phone also raises new ethical questions regarding the trade-off between personal privacy and the public good, and these are questions that Institutional Review Boards (IRBs) [16] and regulators seem as yet poorly equipped to address. Since the absence of both an institutional capacity and an agreed standard of privacy is not slowing the pace of research progress, I would argue that there is at this time a real need not for more *theories* of privacy but for *practices* of data anonymisation that are sensitive to the nature of the investigation and can be applied in real-time to high-volume data sources. To date, applications of spatial and temporal privacy have emerged primarily from work with GPS devices, which have historically used fairly small research populations and are thus not necessarily suited to application in a high-volume data collection system.

In the first part of this work I review the strategies developed in GPS and GIS research as a way of identifying and introducing the key concepts in locational privacy. However, since the algorithms employed in these fields are in many cases too computationally expensive for application ‘as is’ to large volumes of locational data, the second section of this work introduces the concept of Finite State Machines and sets out how they are particularly suited to the specific constraints of cellular network privacy management. Thus the objective of this article is to set out the practical privacy issues associated with the application of computational methodologies to rich behavioural databases, and in particular to data collected from cellular phone networks, and to establish the groundwork for an adaptive, complex system able to respond to these issues in real-time.

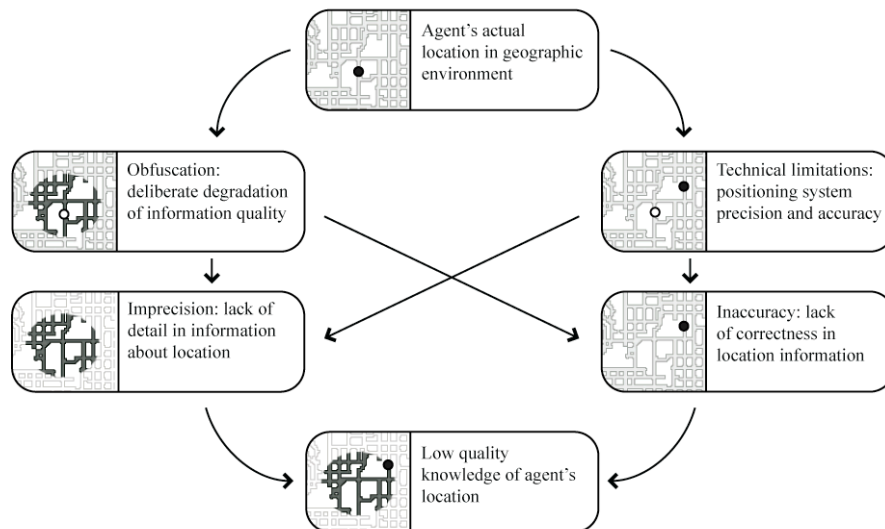
Locational Data & Privacy Issues

Until recently, the cost of hardware and level of sophistication required to follow the digital trails of more than a handful of people at a time meant that the issue of location privacy remained largely theoretical. [3, p.46] But such is the rate of change that we are rapidly moving away from the technical challenges of analysis with ubiquitous devices towards the predominantly social question of what applications are acceptable to the individual and to society as a whole. [15, p.1] A key agent of this transformation is the cellular phone, [2, p.1] which has taken location privacy issues out of the lab and into a largely unprepared world.

There are a host of interesting questions in strategic and transport planning that may be answered with cellular phone data – preferred routes and destinations, mode choice, the relationship between work and non-work trips, face-to-face interaction factors, and more broadly how groups make use of space and time in an urban context – but growing public concern over the use and misuse of personal data may well preclude this type of research. One issue with privacy theory is that there is no ‘silver bullet’ for the public’s privacy concerns because they are largely contextual: few people can give a precise definition of the term [3, p.46] because they typically expect different levels of personal privacy when in a public park or a private home, [15, p.2] and they also respond to the type of information being captured, the entity doing the collection, and the uses to which the data will be put. [2, p.2] And yet, all of these fine distinctions are irrelevant to data collected from a ubiquitous device such as the mobile phone since it can be adapted to nearly any purpose and ignores spatial or temporal niceties.

There is an obvious tension between the quality and quantity of data, and the degree of anonymity retained by those whose details are captured, [24, p.21] since each new scrap of information (date of birth, income, location, etc.) increases the likelihood that an observer – whether a scientist, a bureaucrat, a police officer, or a criminal – is able to identify a ‘person of interest’. In spatiotemporal data there are effectively two axes of *uncertainty* around a single identification: whether the user is who we think they are, and whether they are where we think they are. [10, p.188] This uncertainty can arise in two ways: from limitations inherent in the system itself – how accurate and reliable is the data-collection process itself – and alterations – whether deliberate or inadvertent – of the data that is actually collected. The relationship between these dimensions is nicely captured in Figure 1 below, adapted from Duckham et al. [4, p.50]

Figure 1. Dimensions of Uncertainty



Privacy in Cellular Networks

There are a variety of ways to identify and locate individuals on cellular networks, but they typically involve a trade-off between positioning accuracy and the number of people who can be simultaneously tracked. [26, p.109] So while the police can use sophisticated triangulation algorithms, accurate to within about 15 metres, to track members of a suspected terrorist cell using the mobile network, researchers wishing to understand the commuting habits of millions can only localise people in a fairly rudimentary way using the unique identifier of the cell in the network to which they are connected. The size of cells varies with the density of users, the operating frequency of the radios, and environmental factors: in dense urban areas the cell identifier may enable a user to be localised to within 100m of their true location, but this quickly falls to 600m or more in rural areas where there are far few callers and, hence, fewer cells. [20, p.734]

Since phone companies are understandably wary of sharing data below the cell level, in this article I will take it as a given that an identity attacker only has cell-id data. Figure 2 below shows a fictional cellular network layout in order to help the reader visualise and understand the nature of the problem that will be explored in the rest of this article. Although the reality of antenna coverage areas is known to be more complex, mobile operators often use a Voronoi plot to map out the coverage provided by their antennas. To generate Voronoi cells we can begin with a set of points a through h – where each point is the location of an antenna – and then map out a region X such that every point within X is closer to antenna a than it is to the other antennas b through h . We repeat this process for points b through h and the result is a map that shows which part of the city is most likely to be served by which radio antenna (see Figure 2B).

The path of an individual through the city is registered as series of ‘hand offs’ between adjacent or nearby cells: as the user leaves the area covered by antenna a there is moment where the network registers that calls to this person should now be routed to antenna b instead. If a is at the centre of the Voronoi plot, then we can think of the user’s movement as a series of possible transitions along a network graph: a -to- b , a -to- c , and so on *ad infinitum*. (see Figure 2C) Finally, in this example by recording the transitions $e \rightarrow a$ and $a \rightarrow b$ as they happen, (see Figure 2D) we begin to build the trajectory history of a unique individual.

Figure 2. From Antennas to Trajectories: Map of Antennas (A), Voronoi Plot (B), Possible Handover Transitions (C) & User Path from Handovers (D)

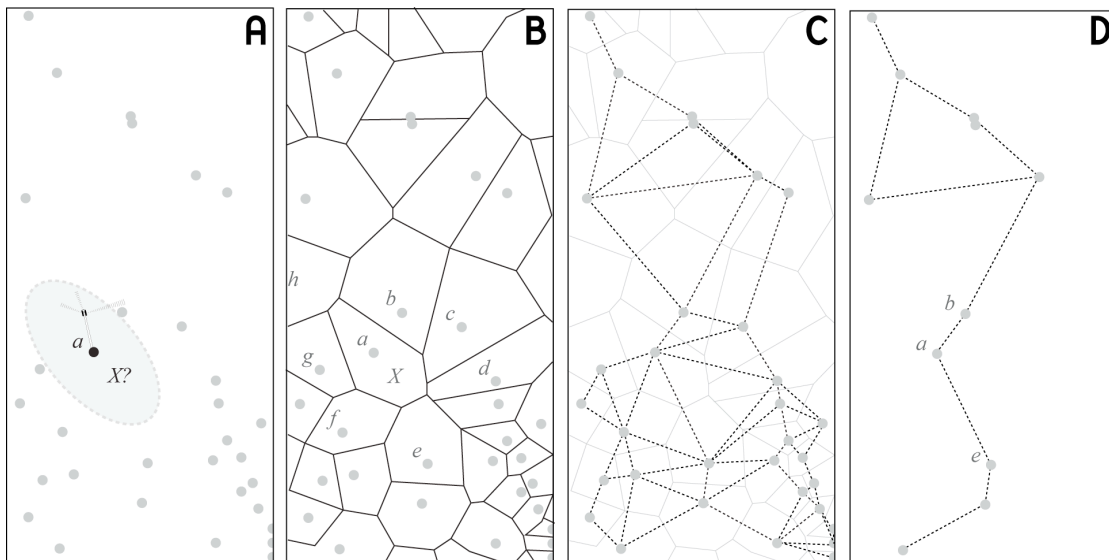


Figure 2 should make clear how, depending on the size of each cell, there will be a degree of imprecision associated with using the cell identifier as a location. For understanding collective behaviour at the urban and regional scales this variability is relatively less important, and so the in-built lack of resolution has privacy benefits for users. However, cellular phones are used very differently from more accurate GPS devices: they are carried everywhere and at all times, so the lack of precision may be counteracted by a continuity of coverage that enables an observer to build a much more complete picture of the user’s behaviour and, consequently, of their identity. Thus it would be naïve – as will be further illustrated below – to rely on the purely technical constraints of the mobile network to ensure personal privacy and we must consider in more detail how indirection and inaccuracy could be deliberately introduced into any collection system.

Concepts in Spatio-Temporal Privacy

As briefly discussed above, the higher accuracy and smaller sample sizes of GPS studies reduce the level of identity uncertainty in the system. Fortunately, in that field the smaller samples also equate to smaller data sets, and so anonymization can be performed well after collection and with less regard for performance. Unfortunately, the scale of cellular network data makes this type of ‘leisurely’ post-processing much more

difficult, and so we must ultimately identify ways of anonymizing or scrambling sensitive information as it is being collected. Nonetheless, there is still a great deal to be learned from GPS-based work about the ways in which locational privacy can be undermined, and we can then adapt these concepts to a more scalable system designed to cope with larger data sets.

Pseudonymity

With a few exceptions, [21, 22] true anonymity makes most kinds of phone-based behavioural research impossible since the connection between an individual and their trajectory or calling history is severed. Instead, many so-called anonymous data sets are actually pseudonymous, in which an arbitrary value – a number or sequence of letters, for instance – is substituted for an identifying name or phone number. [13, p.2]

Pseudonymity is an important mechanism in urban analysis since, for instance, we need a persistent identifier in order to assemble location observations taken at 7am, 7:30am, 8am, 8:30am and 9am into a single morning commute.

Unfortunately, the persistency that allows us to conduct commuting research also gradually erodes individual privacy. As it becomes ever more tightly linked to just one, unique spatiotemporal history, the pseudonym also begins to provide a handle for an identity attacker. [11, p.3] The best answer to this type of attack is to place limits in both space and time upon the duration of a pseudonym. We define a ‘window’ – be it a neighbourhood, a city, a day of week, or number of hours – within which a user retains a pseudonym, but when the limits of the frame are reached then tracking ceases and/or the pseudonym is reset. Windowing makes it possible to track users for research into, say, overall commuting patterns while still making it extremely difficult to expose their identity through an analysis of a long-term travel history.

k-Anonymity & Aggregation

The uniqueness of a mobility history also gives rise to a more subtle issue: the intersection of information taken from more than one data set may still uniquely identify an individual even when any one data set is, on its own, anonymous. [25] For instance, combining commuting behaviour with residence and employment information could unmask a user’s personal travel history even if our travel history were carefully scrubbed of identifying data. One response to this issue, what Sweeney termed *k*-anonymity [24], is effectively a sophisticated form of aggregation: the collection system sets a threshold (*k*) below which individual spatiotemporal data is either not reported at all or is generalised (i.e. made less specific) so that the individual becomes indistinguishable from *k* other users.

In a cellular network context, we can set some minimum number of *k* phone users who must either follow a particular transition (e.g. $a \rightarrow b$, $a \rightarrow g$) or remain within a given cell (e.g. $a \rightarrow a$) in each sampling interval for data on those individuals to be reported (see Figure 2D). Clearly, if there are not enough users to create a crowd that guarantees some level of anonymity then this process could have a serious impact on the types of analysis that are possible. [14, p.9] In GPS research this problem has been acute because of the small sample sizes involved – often less than two hundred people [13, p.3] – but fortunately on telecommunications networks the sample sizes typically reach into the *millions*, suggesting that *k*-anonymity processes may have rather less of a research impact in the final analysis.

Data Perturbation

Another way of introducing uncertainty to a data set is through random ‘noise’ – a user’s location in space and time is adjusted by unpredictable amounts so as to make it more difficult to obtain an accurate location for an individual. Evfimievski, [8] Krumm, [13] and Kwan et al., [14] consider a range of techniques for adding noise to location data; however, the ‘perturbed’ data must be both plausible and misleading: if a path shifts by five kilometres only to revert to a more probable trajectory a few minutes later then we’d have little difficulty in divining an altered data point, but if a user’s location is only shifted by ten metres once in a while then we’ve done little to mislead an attacker. Furthermore, Evfimievski notes several ways in which the distribution of data and the degree of mutual influence between variables might still permit privacy to be

compromised, [8, pp.46-47] and one potentially unanticipated consequence of adding some types of random noise is that the data set may actually become susceptible to noise filtering algorithms. [18, p.18]

Fortunately, there are several aspects of the cellular phone network that reduce the need for algorithmic noise to protect privacy. The first is the nature of the cellular system: areas with few users each cell will tend to cover a larger area, making location less certain, while small cells will correspond to many more simultaneous users, making identity less certainty. The second is that users near the boundary between cells or near transceivers under heavy load (many simultaneous calls) may appear to bounce between cells without having actually moved at all. This effect can give the impression that cells ‘breathe’ inwards and outwards over time, [23] meaning that the Voronoi plot in Figure 2 is better understood as a map of *average* coverage and not *actual* coverage, so there is a less than perfect correspondence between network location and true location. Finally, users travelling at high-speeds across the city may be handed between overlapping ‘macro’ cells to ensure continuity of service, introducing further uncertainty around exact location. [23]

Path Perturbation & Mix Zones

In general, there is an inverse relationship between the density of users and the degree of perturbation required to ensure anonymity. [11, p.10] Consequently, even in entirely anonymous environments if there are too few simultaneous users then tracking algorithms may be able to accurately reassemble a user’s trail using by making reasonable assumptions based on observed direction and speed. [4, pp.50-51] Hoh and Gruteser [11, p.4] suggest that one way to defeat such a tracking algorithms is to use path confusion: each time two users cross paths in time and space there is an opportunity for the attacker to confuse them, especially if we increase the likelihood of this occurring by exchanging the two users’ identifiers.

With smaller data sets it is possible to search for users passing close to one another and adjust their trajectories so that they appear to meet, but this clearly is not possible in a dynamic environment with millions of users. Nonetheless, the method is a particularly appealing one for use in cellular privacy since handovers between cells provide the perfect moment to look users between whom to exchange pseudonyms, and the network cells are sufficiently coarse that we’d expect to have many people in urban areas entering or leaving cells simultaneously and, hence, many opportunities to sow confusion.

As an alternative to path confusion, Beresford and Stajano [3] propose the ‘mix zone’ – a region in which all users either exchange identifiers with one another or switch to new pseudonyms altogether. This method exploits a known tendency for even relatively small populations to cross paths at just a few high-traffic locations such as a corridor or entryway. At the city-scale it is reasonable to expect many more people to enter a mix zone simultaneously – consider Times Square in New York or Oxford Circus in London – suggesting that it would be much more difficult to link users to paths in an urban context. [3, p.54]

Inference Attacks

An attacker can also employ statistical analysis, such as clustering or eigendecomposition, [5] to attack data samples. [13, p.4] By observing and grouping together common behavioural patterns, especially those around work and home life, one can radically increase the likelihood of reidentification. This is not necessarily a direct attack on location *per se*, but by analysing the level of routine in a user’s day we can draw inferences about both the type of user and the likelihood that they will reappear in a particular place at a particular time, significantly reducing the ‘search space’ for an identity-oriented attack. The risk here is then not only in relation to identity but also to prediction and interception.

Krumm is one of the few to evaluate the efficacy of countermeasures against such attacks on GPS data, [13, pp.10-15] and he found that the *minimum* amount of random noise required to reduce to zero the risk of reidentification for his study subjects was more than 2km. Assuming that this is a general case, then no single technique can succeed in verifiably preserving privacy unless the data is corrupted to a degree that renders it unsuitable for many applications. For instance, Krumm notes that adding 2km of noise to a traffic-planning tool would seriously undermine the utility of the analysis. [13, p.14]

Human Error

Finally, most privacy research presumes that secure hardware and software architectures are in place; however, human error and poorly configured systems are often the initial point of failure in privacy breaches. Though not a locational data breach *per se*, the loss in the post of 25 million records containing banking and National Insurance details by Britain's Customs and Revenue (HMRC) Service is only the most egregious of an ongoing series of breaches. [7, 17] If we, as social scientists, wish to start employing this type of sensitive data for public and private research purposes then we must also factor in the increasing risk of disclosure that comes with the wider distribution of this data. We must stop thinking that security breaches only happen to *other* researchers, and consequently I think that there is a clear case for embedding either irreversible anonymization or more sophisticated forms of pseudonymisation even where informed consent has been obtained and the data does not, on the surface, seem particularly sensitive.

Summary

As should by now be clear, the sheer number of mobile phone users and the technical constraints on spatial accuracy in the cellular network provide powerful protection against naïve identity- and privacy-related attacks. However, because the ubiquity of the mobile phone also raises the threshold for what we should consider to be sufficient obfuscation for the purposes of privacy protection – indirectly, we can glean a great deal of knowledge about an individual from monitoring when they wake up, leave the home, travel a preferred route to work, habitually go for lunch, and with whom they spend their free time, and these are avenues for much more sophisticated assaults on personal privacy.

With the benefit of concepts already elaborated in a GPS environment, we have now identified some promising methods for decoupling itinerary and individual. However, the sheer volume of data flowing from the cellular network means that it is effectively impossible to retroactively scrub the data set of identifying information. So we need a mechanism for applying these algorithms or concepts in real-time and without the benefit of a 'big picture' view of network activity. In such a massive behavioural database, we cannot know from minute to minute what events might occur to potentially anonymise or deanonymise any one user.

Finite State Machines & Privacy Management*Overview*

Although they may sound exotic, from a conceptual standpoint Finite State Machines (FSMs) have a great deal in common with agents in geographical agent-based models. In fact, FSMs are embedded in many computer applications, ranging from artificial intelligence research, through video gaming, [19] and on to sophisticated applications in targeted marketing campaigns. [12] FSMs have not been previously applied to privacy management issues, but can be used to address the important issues discussed above, leaving us with an adaptable system with excellent performance characteristics in high-volume data-processing environments.

How do Finite State Machines work?

At their simplest, state machines are described by two mechanisms: states and actions. Actions bring about changes in states, and changes in state can, in turn, induce new actions. For instance, a light switch can be represented as a FSM with two states, on and off, and two actions 'close the relay' (i.e. turn the light on) and 'open the relay' (i.e. turn the light off). When the switch is moved to the 'on' state then the 'close the relay' action occurs, and when moved to 'off' then the 'open the relay' action happens automatically. By adding more states, more actions, and more types of actions we can model systems of immense complexity with fairly few, simple 'rules'.

Actions have four distinct sub-classes: entry, exit, input, and transition. Input actions enable external events to bring about changes of state. Entry actions are performed when a machine enters a new state, exit actions occur on leaving the current state, and transition actions occur while the machine is moving from one state to another. So in a more complex environment, we could use state machines to manage a marketing campaign: first-time buyers (an action) of a product become customers (a state) of the marketing firm; the input

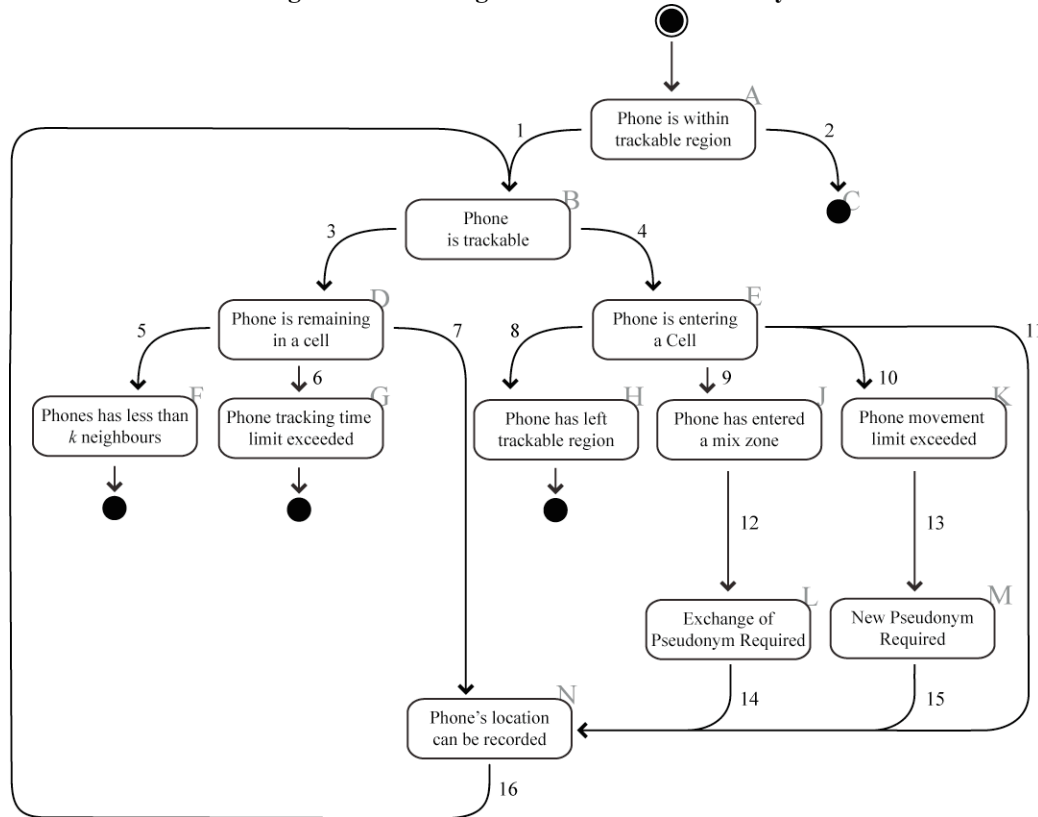
action of calling a support centre for help with the new purchase can trigger the customer’s reassignment to an in-need-of-support state whose entry action is a follow-up assurance call or a ‘getting the most of your new product’ promotion. If the customer then indicates satisfaction with their purchase they can be moved back into the default customer state and become ‘eligible’ for other promotions.

Applying FSMs to Cellular Network Privacy

As quantitative social scientists doing locational research we want to record two pieces of information about users connected to the cellular network: their location at time t_n as captured by a cell id, and any changes in location at time t_{n+1} captured as the handover of the user from one cell to another. If we discard for a moment the particularities of geography, then we can represent the cellular system as a network of nodes and links (e.g. $e \rightarrow a$ and $a \rightarrow b$) as illustrated in Figure 2C and Figure 2D. This representation is also amenable to graph analysis of the sort performed by González et al. [9] to study human mobility patterns, and would take the more static analyses performed by groups such as Space/Syntax in a substantively new direction.

However, the focus of this article is privacy and so this consideration should drive the majority of actions and states in our data management process. Figure 3 sets out one possible implementation graphically, while Table 1 presents the underlying logic as a state transition table with pseudo-code. In our simple model, a user enters an initial state *A* when their phone connects to a cell within our track-able region. Transitions 1 and 2 represent two possible exit actions that are effectively setting the sampling rate for the general population of track-able users. Transition 2 leads to a terminal condition since no further actions are performed on the handset, and these can arise anywhere in the system of state machines. In this example there are three other terminal states where users are dropped from the processing model.

Figure 3. Modelling Cellular Network Privacy



The FSM approach enables us to implement a range of privacy-preserving techniques with much less computational sophistication than is required when just one algorithm is applied post-collection. For instance, the transitions leading to states *G* and *K* implement windowing, the action leading to state *F* implements *k*-anonymity, and the one to state *J* implements mix-zones in specified cells. We could compound the difficulty of

a privacy attack by applying random modifiers to transitions 6 and 10 so that the windows for each handset vary unpredictably. The FSM system can even be quickly rewired such that state *D* is connected to state *M* by a transition that causes some percentage of users to receive a new pseudonym without having reached the hard limit set by the window. So instead of a single algorithm that functions well under a particular set of data collection conditions, we are creating a system that is arbitrarily complex so that it is impossible to know which, if any, attack or set of attacks would have a higher probability of success.

The benefits of FSMs are thus substantial: first, because they are computationally simple they can be used in data-rich/high-throughput environments where real-time performance is critical; second, they provide a straightforward way of mapping a set of desired privacy and performance objectives on to a fully-fledged application; and third, although it is easy to describe and model the ‘states’ through which an individual user might pass it is impossible to predict their actual behaviour or the behaviour of the system as a whole. As with agent-based models, a set of simple rules – if *x* event occurs then take *y* action – can give rise to unpredictable behaviours on the large scale. And because each transition can be tuned with a separate probability we can also deliver a range of privacy characteristics that can be adjusted over time.

Table 1. Model State Transition Table

| Transition | State | | | | | | | | | | | | | | |
|------------|-------|---|---|---|---|---|---|---|---|---|---|---|--|--|---|
| | A | B | C | D | E | F | G | J | K | L | M | N | | | |
| 1 | B | | | | | | | | | | | | | | Sampling and load management: |
| 2 | C | | | | | | | | | | | | | | goto B if $r < 0.5$ else C |
| 3 | D | | | | | | | | | | | | | | Same cell in last interval: |
| 4 | E | | | | | | | | | | | | | | goto D if $loc(U, t) == loc(U, t_{-1})$ else E |
| 5 | | | F | | | | | | | | | | | | k-anonymity: goto F if $count(Loc_i, t) < k$ |
| 6 | | | G | | | | | | | | | | | | Temporal window: goto G if $U(t - t_0) > c$ |
| 7 | | | N | | | | | | | | | | | | Default: goto N |
| 8 | | | | H | | | | | | | | | | | Left region: goto H if $loc(U, t) \notin Loc$ |
| 9 | | | | | J | | | | | | | | | | Mix zones: goto J if $loc(U, t) \in Mix$ |
| 10 | | | | | K | | | | | | | | | | Spatial window: goto K if $count(Loc_u) > s$ |
| 11 | | | | | N | | | | | | | | | | Default: goto N |
| 12 | | | | | | | | L | | | | | | | Exchange id: $ex(U_1, U_2)$ if $loc(U_1, t) == loc(U_2, t)$ |
| 13 | | | | | | | | | M | | | | | | New id: $U_1 = id(s * IMSI)$ |
| 14 | | | | | | | | | | N | | | | | Default: goto N |
| 15 | | | | | | | | | | | N | | | | Default: goto N |
| 16 | | | | | | | | | | | | B | | | Default: goto B |
| 17 | | A | | | | A | A | | | | | | | | Optional: goto A |

r is a random number between 0 and 1

s is a salt whose value varies over time

Loc is the set of trackable locations, *Loc_i* is a location within set *Loc*, *Loc_u* is the set of locations at which user *U* has been found

Mix is the set of mix zones, Mix_i is a mix zone within set Mix

T is the most recent observation time, t_{-1} the preceding observation time, t_0 the first observation time

$Loc(U, t)$ is the location data point of user U at time t

$Id()$ is a pseudonym-generating function

$Ex()$ is a pseudonym-exchanging function

Conclusions & Next Steps

In large behavioural data sets it is usually our habits that give us away, and since we are very much creatures of habit [5,9] the risk of extended and omnipresent ‘monitoring’ as a vector for identity-related attacks is substantial. Yet accurate and extensive locational data can also be used to enhance public welfare by improving our understanding of how cities function in space and time and, consequently, our ability to deliver both public and private goods in a adaptable and sustainable manner. To date, the public response to geoprivacy issues has been surprisingly muted, but ongoing high-profile breaches suggest that data protection and privacy will remain in the public eye for the foreseeable future, and objections to large scale data collection and sharing by private entities is likely to increase rapidly.

Accordingly, there is a strong incentive for researchers and commercial organisations to collaborate on developing effective anonymization systems since failure to do so may well preclude the collection of any data at all. [16] By enabling us to introduce several types of unpredictability simultaneously, Finite State Machines enable us to respond to the privacy concerns of individuals while still supporting the development of the cellular network as a platform for urban analysis. In effect, we can tune the state machines so that data sets that are demonstrably corrupted from the standpoint of individual identity, [13, p.11] but nonetheless preserve the overall characteristics and distributions of urban activity. Additional work is required to determine the appropriate thresholds and types of noise that would deliver the optimal mix of privacy and statistical validity, but by enabling the fine-grained analysis of behaviour in space and time this approach advances our understanding of the city as an integrated whole without violating the expected privacy rights of users.

Acknowledgements

I wish to thank Peter Hall and the Balzan Foundation, whose generous support for early-stage researchers has made this work possible, as well as Francesco Calabrese at MIT’s SENSEable City Laboratory for his insightful comments. Finally, I would like to recognise Richard Vermillion, Chief Technology Office at Fulcrum Analytics, for introducing me to the concept of Finite State Machines and for demonstrating the wide range of problems to which they can be applied – technical or logical oversights in this paper are no reflection on his capabilities.

References

1. R. Ahas & Ü. Mark, “Location Based Services – New Challenges for Planning and Public Administration?” *Futures*, 37:6 (2005) 547-561.
2. L. Barkhuus & A. Dey, “Location-based services for mobile telephony: a study of users’ privacy concerns,” paper presented at *INTERACT2003: 9th IFIP TC13 International Conference on Human-Computer Interaction* (Zürich, Switzerland, September 1-5, 2003).
3. A. Beresford & F. Stajano, “Location Privacy in Pervasive Computing,” *IEEE Pervasive Computing* 2:1 (2003) 46-55.
4. M. Duckham, L. Kulik & A. Birtley, “A Spatiotemporal Model of Strategies and Counter-Strategies for Location Privacy Protection,” paper presented at *4th International GIScience Conference* (Münster, Germany, September 20-23, 2006).
5. N. Eagle & A. Pentland, “Eigenbehaviors: Identifying Structure in Routine,” *Behavioral Ecology and Sociobiology*, 63:7 (2009) 1057-1066.
6. N. Eagle, “Behavioral Inference across Cultures: Using Telephones as a Cultural Lens,” *IEEE Intelligent Systems*, 23:4 (2008) 62-64.
7. Economist, “Identity Parade,” *The Economist* (February 14, 2008).
8. A. Evfimievski, “Randomization in Privacy Preserving Data Mining,” *ACM Special Interest Group on Knowledge Discovery and Data Mining Explorations*, 4:2 (2002) 43-48.
9. M. González, C. Hidalgo & A.L. Barabasi, “Understanding individual human mobility patterns,” *Nature*, 453 (2008) 779-782.
10. M. Gruteser & B. Hoh, “On the Anonymity of Periodic Location Samples,” paper presented at *Proceedings of the Second International Conference on Security in Pervasive Computing* (Boppard, Germany, April 6-8 2005).
11. B. Hoh & Gruteser M., “Protecting Location Privacy Through Path Confusion,” paper presented at *First International Conference on Security and Privacy for Emerging Areas in Communications Networks* (Athens, Greece, September 5-9, 2005).
12. D. King, “Using Finite State Machines to Manage Customer Relations,” *Multichannel Merchant* (September 4 2007) <http://multichannelmerchant.com/crosschannel/lists/state_machines_09042007/index.html> Accessed February 3, 2008.
13. J. Krumm, “Inference Attacks on Location Tracks,” paper presented at *Fifth International Conference on Pervasive Computing* (Toronto, Canada, May 13-16, 2007).
14. M.-P. Kwan, I. Casas & B. Schmitz, “Protection of Geoprivacy and Accuracy of Spatial Information: How Effective are Geographical Masks?” *Cartographica*, 39:2 (2004) 15-28.
15. M. Langheinrich, “Privacy Invasions in Ubiquitous Computing,” *Ubicomp 2002 Privacy Workshop* (2002) <<http://guir.berkeley.edu/pubs/ubicomp2002/privacyworkshop/papers/uc2002-pws.pdf>> Accessed February 8, 2008.
16. D. Lazer, A. Pentland, L. Adamic, A.L. Barabasi, D. Brewer, N. Christakis, N. Contractor, J. Fowler, M. Gutmann, T. Jebara, G. King, M. Macy, D. Roy & M. Van Alstyne, “Computational Social Science,” *Science*, 323 (2009) 721-723.
17. J. Leyden, “Information security breaches quadrupled in 2007,” *The Register* (2008) <http://www.theregister.co.uk/2008/01/02/data_breaches_skyrocket/> Accessed February 22, 2008.
18. B. Malin, “Betrayed By My Shadow: Learning Data Identity via Trail Matching,” *Journal of Privacy Technology* (2005) 20050609001.
19. N. Meyer, “Finite State Machine Tutorial,” *Generation5* (2003) <http://www.generation5.org/content/2003/FSM_Tutorial.asp> Accessed February 20, 2008.
20. C. Ratti, R. Pulselli & S. Williams, “Mobile Landscapes: using location data from cell phones for urban analysis,” *Environment & Planning B: Planning and Design*, 33:5 (2006) 727-748.

21. J. Reades, F. Calabrese, A. Sevtsuk & C. Ratti, “Cellular Census: Explorations in Urban Data Collection,” *IEEE Pervasive Computing*, 6:3 (2007) 30-38.
22. J. Reades, F. Calabrese & C. Ratti, “Eigenplaces: analysing cities using the space-time structure of the mobile phone network,” *Environment & Planning B: Planning and Design*, forthcoming.
23. A. Sevtsuk, personal communication (April 5, 2007).
24. L. Sweeney, “Information Explosion,” in L. Zayatz, P. Doyle, J. Theeuwes and J. Lane, eds., *Confidentiality, Disclosure, and Data Access: Theory and Practical Applications for Statistical Agencies* (Washington: Urban Institute, 2001).
25. L. Sweeney, “Achieving k -anonymity privacy protection using generalization and suppression,” *International Journal on Uncertainty, Fuzziness and Knowledge-based Services* 10:5 (2002) 571-588.
26. Y. Zhao, “Standardization of Mobile Phone Positioning for 3G Systems,” *IEEE Communications Magazine* 40:7 (2002) 108-116.